Eğitim ve Bilim

Cilt 40 (2015) Sayı 177 285-296

Performansa Dayalı Durum Belirlemede Güvenirliğin Genellenebilirlik Kuramında Farklı Desenlerle İncelenmesi *

Serap Büyükkıdık [1], Duygu Anıl [2]

Öz

Bu araştırmada, genellenebilirlik kuramında farklı desenler kullanılarak yapılan güvenirlik sınamasıyla, yüzey (facet) sayısı değişiminin güvenirliği nasıl etkilediği incelenmiştir. Araştırma verileri 2011- 2012 Bahar döneminde Kütahya ilinde bir ortaokulda eğitim gören 132 6-7-8. sınıf öğrencisinin rutin olmayan problem çözümüne yönelik sergiledikleri performansların dört puanlayıcı tarafından, göreve özgü analitik ve bütünsel dereceli puanlama anahtarı kullanarak puanlanmasından elde edilmiştir. Araştırmada, bireyin ölçme objesi; görev, puanlayıcı (b:birey, görev:g, p:puanlayıcı) ve dereceli puanlama anahtarı (a:anahtar) çeşidinin değişkenlik kaynağı olarak alındığı bxgxp ve bxgxpxa tümüyle çapraz desenleri kullanılmıştır. Araştırma sonucunda, genellenebilirlik kuramında kullanılan desenlerin G ve phi katsayılarını etkilediği, değişkenlik kaynağı sayısı arttıkça ölçmenin amacı bireyin toplam varyansı açıklama yüzdesinin azaldığı görülmüştür. Ayrıca dereceli puanlama anahtarı türünün de ölçümlerin güvenirliğini etkileyebileceği, analitik dereceli puanlama anahtarından elde edilen puanların göreceli olarak bütünsel dereceli puanlama anahtarından elde edilenlere oranla daha yüksek güvenirliğe sahip olduğu bulunmuştur.

Anahtar Kelimeler

Performansa dayalı durum belirleme

Genellenebilirlik kuramı

Puanlayıcılar arası güvenirlik

Analitik dereceli puanlama anahtarı

Bütünsel dereceli puanlama anahtarı

Makale Hakkında

Gönderim Tarihi: 04.12.2012

Kabul Tarihi: 23.10.2014

Elektronik Yayın Tarihi: 15.02.2015

DOI: 10.15390/EB.2015.2454

Giriş

Değerlendirmenin niteliği ne olursa olsun, tek doğru yerine esnek yanıtlama imkanı olan yanıtlar amaçlanılmalı, bu yüzden makinalar tarafından yapılan objektif puanlamaların yerine, öğrencilerin kendi yanıtlarını oluşturmalarını gerektiren, insan kanısına dayalı puanlamalar kullanılmaktadır (Linn ve Miller, 2004: s. 6). Üst düzey davranışların ölçülmesinde okullarda objektif ve hızlı puanlama yapılabilmesi, daha fazla soruya yer vererek ölçülmek istenen kapsamı daha iyi yansıtabilmesi nedeniyle sıklıkla kullanılan çoktan seçmeli testler yetersiz kalabilmektedir. Performansa dayalı durum belirleme çoktan seçmeli testlerin bu sınırlılığını gidermede kullanılabilecek yöntemlerden biridir. Performansa dayalı durum belirlemede puanlayıcıların

* Bu çalışma "Büyükkıdık, S. (2012). Problem Çözme Becerisinin Değerlendirilmesinde Puanlayıcılar Arası Güvenirliğin Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Karşılaştırılması." başlıklı yüksek lisans tezinin bir kısmından oluşturulmuştur.

[1] Gazi Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri, Ölçme ve Değerlendirme Bölümü, Türkiye, sbuyukkidik@gmail.com

[2] Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri, Ölçme ve Değerlendirme Bölümü, Türkiye, aduygu@hacettepe.edu.tr

kanaatleri devreye girdiğinden puanlama yöntemi ve puanlayıcı güvenirliği önem taşımaktadır. Güvenirliğin sınanmasında ise çeşitli kuramlar ve uygulamaları bulunmaktadır. Kuramların ve kuramların uygulanmasındaki desenlerin farklılaşması elde edilen güvenirlik katsayılarının farklılaşmasına, uygulamalardan farklı bilgiler edinilmesine sebep olabilmektedir.

Bu çerçevede performansa dayalı durum belirlemede puanlama yöntemi olarak kullanılan analitik ve bütünsel puanlama anahtarlarıyla elde edilen puanların güvenirliği genellenebilirlik kuramında farklı desenlerin kullanılmasıyla elde edilen bilgiler doğrultusunda karşılaştırılacaktır.

Performansa Dayalı Durum Belirleme

Performansa dayalı durum belirleme, öğrencilerin sahip olduğu "özel bilgileri, kavramları ya da becerileri gerçek yaşam içeriği ya da koşulunu" yansıtan görevlerle gösterdikleri bir durum belirleme yöntemi olarak tanımlanır (American Educational Research Association, American Psychological Association ve National Council on Measurement in Education, 1999). Performansa dayalı durum belirleme, öğrencilerin bir aktivite yerine getirmesini (örneğin; model oluşturma) ya da orjinal bir yanıt oluşturmasını (örneğin; bir matematik probleminin çözümünü açıklamak); üst düzey düşünmeyi ve problem çözmeyi ölçen; öğrencilerin gerçek yaşam durumlarıyla ilişkili problem çözme uygulamasını gerektiren, çoklu çözüm yolu ve yanıtları olan, önceki bilgilerine ulaşmalarını sağlayan ve birkaç dakikadan birkaç gün ve daha fazlasına yayılabilen zaman dilimi gerektiren uygulamalar olarak tanımlanabilir (Aschbacher, 1991; Baron, 1991; Herman, Aschbacher ve Winters, 1992; Madaus ve O'Dwyer, 1999; Stiggins, 1987; Akt: Lane ve Stone, 2006: s. 387).

Dereceli Puanlama Anahtarları

Dereceli puanlama anahtarı (rubric) sözcüğünün orjinal anlamı, öğrenci işlerini puanlamakla çok az ilgilidir. Oxford İngilizce Sözlüğü rubriği, 15 yüzyılın ortalarında bir kitabın farklı bölümlerinin başlığı olarak tanımlar. Bu sözcüğün kökeni dinsel alanyazını dikkatli bir şekilde yeniden yapılandıran hristiyan rahiplerin işlerine dayanıyordu, başlangıçta kopyalanan kitabın her bir ana bölümü kırmızı büyük harflerle başlıyordu. Latincede "ruber" kelimesi kırmızı olduğundan, rubric bir kitabın ana ayrımları için önem arz eden başlıklar anlamına geldi. Son yirmi-otuz yıl öncesinde, rubrik eğitimciler arasında yeni anlamını almaya başladı (Popham, 1997: s. 72). Goodrich (1997: s.14)'e göre, dereceli puanlama anahtarı (rubric), bir iş için ölçütleri ya da "ne ölçülüyor'u'' (Örneğin; yazma becerisi için: amaç, organizasyon, ayrıntı, ses ve yazmanın sıklıkla kullanılan diğer parçalarının mekanizması) listeleyen, hem de her bir ölçüt için o işteki başarının yetkinden, zayıfa doğru geçişlerini gösteren bir puanlama aracıdır. Nerede ve ne zaman dereceli puanlama anahtarının kullanılacağı sınıf seviyesi ya da konuya bağlı değildir, ancak durum belirlemenin amacına bağlıdır (Moskal, 2000). Dereceli puanlama anahtarı da kendi arasında performansı parçalara ayırarak değerlendiren analitik dereceli puanlama anahtarı ve performansın bütününe odaklanan bütünsel dereceli puanlama anahtarı olmak üzere ikiye ayrılmaktadır (Mertler, 2001).

Bütünsel Dereceli Puanlama Anahtarı (Holistic Rubric)

Holistik kelimesi yunanca "holos" kelimesinin kökünden türemiştir ve "bütün, tam" anlamındadır. Holistik yöntemler "genel izlenim" ya da "bütün izlenim" olarak da bilinmektedirler (Priestley, 1982: s. 203). Bütüncül (bütünsel) dereceli puanlama anahtarları, öğretmenlerin genel süreci ürünü bütün olarak parçaları dikkate almadan puanlamasını içerir (Nitko, 2001; Akt. Mertler, 2001). Performansın ya da ortaya çıkan ürünün bileşenlerini yargılamadan ürünü ya da performansı bir bütün olarak değerlendirmeyi gerektirir (Moskal, 2000).

Analitik Dereceli Puanlama Anahtarı (Analytic Rubric)

Analitik dereceli puanlama anahtarları önce performans veya ürünün parçalarının ayrı ayrı puanlanmasını sonra da bu bireysel puanları toplayarak hesaplanmasını içerir (Moskal, 2000).

Genellenebilirlik Kuramı

Genellenebilirlik kuramı, klasik test kuramının (KTK) günümüzde hala yaygın kullanılan gerçek puan modelinin sınırlılıklarına olan tepkiler üzerine Cronbach, Gleser, Nanda ve Rajaratnam tarafından (1963-1972) ortaya atılmış; Shavelson and Webb (1991), Brennan (1992) ve son olarak Brennan (2001a)'nın çalışmaları doğrultusunda geliştirilmiş; davranış ölçümlerinde güvenirliğin değerlendirilmesini, güvenilir gözlemlerin G (genellenebilirlik) ve K (karar) çalışmalarıyla tasarlanmasını, araştırılmasını ve gözlenen puanlardaki tutarsızlık kaynaklarının miktarının tek bir katsayı ile belirlenmesini sağlayan, temelinde varyans analizine (ANOVA) dayalı olan istatistiksel bir kuramdır (Shavelson ve Webb, 1991; Brennan, 2001a, 2001b).

Genellenebilirlik Çalışması

G kuramında, genellenebilirlik (generalizability) katsayısı olarak adlandırılan bir katsayı hesaplanır. Bu katsayı KTK'deki güvenirlik katsayısına benzemekle birlikte güvenirlik kavramını yeniden yorumlamaz. G kuramı, güvenirlik ve geçerlik arasındaki geleneksel farkın güvenilir gözlemler düzenlemekle nasıl ortadan kaldırılabileceğini de gösterir. G kuramında bir evren, onun değişkenlik kaynakları ve gözlemlerin koşulları geçerlik kavramının geleneksel alanında açıklanan yapıdan tanımlanır. G kuramı, gözlemlerin örtülü yapısı (kabul edilebilir gözlemler evreni) hakkındaki vardamaları doğrulukla göstermeyi sağlarsa, gözlemleri güvenilir olarak tanımlar (Shavelson ve Webb, 1991). G çalışmasının amacı, kabul edilebilir gözlemler evreniyle varyans bileşenlerinin kestirimini elde etmektir (Brennan, 2001a: s. 8) .

Karar Çalışması

Genellenebilirlik kuramı, karar (K) çalışmasını genellenebilirlik (G) çalışmasından ayırır. K çalışması, G çalışmasından elde edilen bilgileri kullanarak belli bir amaçla yapılan bir ölçmedeki hataları en aza indirmenin yollarını araştırmak için düzenlenir (Crocker ve AIgina 1986; Shavelson ve Webb 1991; Brennan, 2001a).

Belki de K çalışmasının en önemli katkısı karar vericilerin genellemek istediği belirli ölçme yöntemlerinin sonuçlarına dayalı olan genelleme evreninin belirlenmesidir (Brennan, 2001a: s. 8).

Genellenebilirlik Kuramında değişkenlik kaynağının sayısına bağlı olarak desenin oluşturulmasının yanı sıra, çaprazlanmış (crossed) ya da yuvalanmış (nested) olmak üzere iki türden desen vardır. Ölçmedeki değişkenlik kaynaklarının bütün koşulları diğer bir değişkenlik kaynağının bütün koşullarını etkiliyorsa çaprazlanmıştır ve değişkenlik kaynakları arasına "x" işareti konularak gösterilir. Bir değişkenlik kaynağının bazı koşulları, diğer bir değişkenlik kaynağının bazı koşullarınca gözlemleniyorsa yuvalanmıştır ve değişkenlik kaynakları arasına ":" konularak gösterilir (Shavelson ve Webb, 1991; Brennan, 2001a; Mushquash ve O'Connor, 2006). Ölçmenin hedefi bu çalışmada bireyler olduğundan birey (b) genellikle "facet (yüzey)" yani "ölçme hatasının olası kaynağı" olarak adlandırılmaz.

Araştırmanın Amacı

Bu araştırmada genellenebilirlik kuramında farklı desen uygulamasının güvenirliğe etkisinin incelenmesi amaçlanmıştır. Bu amaçla aşağıdaki sorulara yanıt aranmıştır.

1. Analitik ve bütünsel dereceli puanlama anahtarından elde edilen puanlarda birey (b) ölçme objesi ile, görev (g) ve puanlayıcı (p) değişkenlik kaynaklarının çapraz tasarlandığı b x g x p ve anahtar (a) değişkenlik kaynağının ele alındığı b x g x p x a desenlerinin G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri nasıldır?
2. Analitik ve bütünsel dereceli puanlama anahtarından elde edilen puanlarda b x g x p ve anahtar değişenlik kaynağının ele alındığı b x g x p x a desenlerinin puanlayıcı ve görev sayılarının artırılıp, azaltılmasıyla yapılan karar çalışması (K) sonucunda elde edilen G ve phi katsayıları nasıldır?
3. Güvenirlik sınanmasında kullanılan b x g x p ve b x g x p x a desenlerinden elde edilen bulgular farklılaşmakta mıdır?

Araştırmanın Önemi

Alanyazında elde edilen ölçümlerin güvenirliği ve genellenebilirliğinin genellenebilirlik kuramı ile sınanmasında birçok çalışma karşımıza çıkmaktadır (Al-Mahroos, 2009; Arce-Ferrer ve Castillo, 2007; Atılgan, 2004; Christ et. al., 2010; Deliceoğlu, 2009; Eser, 2011; Güler, 2008; 2009; 2011; Hoyt ve Melby, 1999; Jarjoura et al, 2004; Kaya, 2011; Kozaki, 2004; Nalbantoğlu, 2009; 2012; Öztürk, 2011; Taşdelen ve diğerleri, 2010; Tindal et. al, 2010; Van Hooft et. al., 2006; Yelboğa, 2007; 2012). Tüm bu çalışmalar incelendiğinde farklı dereceli puanlama anahtarlarının kullanılıp, dereceli puanlama anahtarının da değişkenlik kaynağı olarak alındığı tümüyle çaprazlanmış desenlerin genellenebilirlik kuramında uygulanıp, bulguların karşılaştırıldığı benzer bir araştırmaya rastlanmamıştır. Bu çalışma ile birey, görev ve puanlayıcı değişkenlik kaynaklarının tümüyle çaprazlandığı iki yüzeyli desen ile bu değişkenlik kaynaklarının yanında dereceli puanlama anahtarının da ele alındığı üç yüzeyli tümüyle çaprazlanmış desenden elde edilen bulgular karşılaştırılarak, alanyazına katkı sağlanabileceği düşünülmektedir.

Yöntem

Araştırmanın Türü

Araştırma, genellenebilirlik kuramında farklı desen uygulamasının güvenirliğe etkisini incelemesiyle, var olan durumu ortaya koyduğundan betimsel bir araştırmadır.

Verilerin Toplanması

Araştırmanın verileri, 2011- 2012 eğitim-öğretim yılında Kütahya iline bağlı merkez ilköğretim okulunda öğrenim gören 132 ilköğretim 6.,7.,8. sınıf öğrencisinin rutin olmayan problem çözme becerisine yönelik hazırlanan iki performans görevinde sergiledikleri performansların ilköğretim matematik öğretmeni dört puanlayıcı tarafından yanıt tanıma kodları ve analitik dereceli puanlama anahtarıyla ve ardından on gün arayla bütünsel dereceli puanlama anahtarıyla puanlanması ile toplanmıştır.

Polya (1973: 171)'in de belirttiği gibi genelde bir problem önceden çözülmüş genel bir probleme özel veriler yerleştirerek ya da hiçbir yenilik yaratmaksızın iyice bilinen bir örneği adım adım izleyerek çözülebiliyorsa rutin bir problemdir. Araştırmada kullanılan performans görevleri araştırmacı tarafından alanyazın taranarak oluşturulmuş tek doğru çözüm yolu ve yanıtı olmayan rutin olmayan problemleri içermektedir. Araştırmada kullanılan performans görevlerinin öğrenci seviyesine, problem çözme becerisini ölçmeye uygun olup olmadığına aralarında ilköğretim matematik öğretmenleri, matematik eğitimi ve ölçme ve değerlendirme uzmanlarının yer aldığı on kişinin görüşleri alınarak karar verilmiştir. Aynı şekilde dereceli puanlama anahtarları için de bu uzmanlardan ve uzman görüşü alınmıştır. Performans görevlerinin ve dereceli puanlama anahtarlarının Türkçe'ye uygunluğunun belirlenmesi için ise bir Türk dili ve edebiyatı öğretmeninin görüşüne başvurulmuştur.

Puanlayıcılar puanlamanın nasıl olması gerektiğine dair araştırmacı tarafından verilen eğitimin ardından, öncelikle davranış tanıma kodları kullanarak her bir kağıdı içerik analiziyle incelemişlerdir. Ardından puanlayıcılar problemi anlama, çözüm yolları belirleme ve uygulama, çözümü belirtme ölçütlerini ve beş düzeyi barındıran göreve özgü analitik puanlama anahtarıyla her bir performans için puanlamalarını arka arkaya gerçekleştirmişlerdir. Unutma zamanı on gün geçtikten sonra bütünsel dereceli puanlama anahtarıyla aynı süreç gerçekleştirilmiştir.

Verilerin Analizi

Genellenebilirlik kuramında analitik ve bütünsel puanlama anahtarı için b x g x p ve anahtar çeşidinin değişkenlik kaynağı olarak alındığı b x g x p x a tümüyle çaprazlanmış desenleri kullanılarak güvenirlik analizi yapılmıştır. Araştımanın birinci aşamasında, genellenebilirlik kuramında her iki puanlama anahtarı için G çalışması yürütülerek ana ve ortak etkiler için b x g x p deseninde varyans değerleri kestirilmiştir. Araştırmanın ikinci aşamasında genellenebilirlik kuramı kapsamında b x g x p x a deseninde G çalışması yürütülerek ana ve ortak etkiler için varyans değerleri analizleri yapılmıştır. Üçüncü aşamada ise aynı desen için karar çalışması yürütülerek puanlayıcı ve görev sayısının bir arttırılıp azaltılması durumunda G ve phi katsayısı kestirimlerine yer verilmiştir. Son aşamada ise kullanılır her iki desenden elde edilen değerlerin güvenirliği nasıl etkilediği incelenmiştir. Verilerin analizinde EduG 6.0 e programından yararlanılmıştır.

Bulgular

G Çalışması Bulguları

İlköğretim ikinci kademe 132 öğrencinin iki göreve gösterdikleri performansların analitik ve bütünsel dereceli puanlama anahtarı ile dört puanlayıcı tarafından puanlanmasından elde edilen puanlara G kuramında b x g x p ve b x g x p x a desenleri ile genellenebilirlik çalışması yapılmıştır. Her bir değişkenlik kaynağının kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri Tablo 1'de verilmiştir.

Tablo 1. b x g x p ve b x g x p x a Desenlerine Ait G Çalışması Sonucunda Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

desen		**Bxgxp** *Analitik dereceli puanlama*		**Bxgxp** *Bütünsel dereceli puanlama*		**bxgxpxa**	
Varyans Kaynağı	**Sd**	**Varyans σ^2**	**%**	**Varyans σ^2**	**%**	**Varyans σ^2**	**%**
b	131	15.423	80.6	14.836	76.0	14.149	72.5
g	1	0.107	0.6	0.243	1.2	0.084	0.4
p	3	0.078	0.4	0.310	1.6	0.274	1.4
a	1	-	-	-	-	0.034	0.2
bg	131	0.004	0.0	0.000	0.0	0.000	0.0
bp	393	2.344	12.2	2.191	11.2	0.074	0.4
ba	131	-	-	-	-	2.946	15.1
gp	3	0.025	0.1	0.042	0.2	0.000	0.0
ga	1	-	-	-	-	0.016	0.1
pa	3	-	-	-	-	0.000	0.0
bgp	393	1.154	6.1	1.901	9.8	0.017	0.1
bga	131	-	-	-	-	0.388	2.0
bpa	393	-	-	-	-	0.416	2.1
gpa	3	-	-	-	-	0.042	0.2
bgpa,e	393	-	-	-	-	1.072	5.5
Toplam	2111		100		100		100

b: birey, g: görev, p: puanlayıcı, a: dereceli puanlama anahtarı

Tablo 1'de verilen analitik dereceli puanlama anahtarı kullanılarak elde edilen verilerde b x g x p desenine ait G çalışması sonucunda kestirilen varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde, en büyük varyans bileşeninin problem çözme becerileri bakımından farklılık gösteren birey (b) ana etkisinin olduğu σ_b^2 (15.423) varyans bileşeni değeri ve % 80.6 toplam varyansı açıklama yüzdesiyle görülmektedir. Bu performansa dayalı durum belirleme uygulamasında amaçlanan bireylerden kaynaklanan beceri farklılıklarının yansıtılabildiğinin ve problem çözme becerisi bakımından değerlendirme yapılan grubun heterojen özellikler gösterdiğinin göstergesi olabilir. Birey x görev ortak etkileşimi ölçülmek istenmeyen görev etkisinin, ölçülmek istenen birey etkisini etkilemesi sonucu ortaya çıkan değişkenlik kaynağıdır. Bireylerin bir görevden diğerine durumlarının değişmediği σ_{bg}^2(0.004) varyans oranı ve % 0.0 toplam varyansı açıklama yüzdesiyle görülmektedir. Puanlayıcıların bireylerin performansını değerlendirirken birbirlerine göre daha cömert (yumuşak) olmadığını σ_p^2(0.078) varyans bileşen değeri ve % 0.4 toplam varyansı açıklama yüzdesiyle görülmektedir. Farklı puanlayıcıların benzer puanlamalar yaptığı anlaşılmaktadır. Birey görev puanlayıcı etkileşimi (bgp) $\sigma_{bgp,e}^2$ ölçme hatasından kaynaklı artık değişkenlik kaynağıdır. Tablo 1 incelendiğinde; ölçmenin nesnesi birey ve b x p ortak etkileşiminden sonra tesadüfi hata olarak da adlandırılan bu değişkenlik kaynağının en büyük değişkenlik kaynağı olduğu $\sigma_{bgp,e}^2$(1.154) varyans bileşen değeri ve % 6.1 toplam varyansı açıklama yüzdesiyle anlaşılmaktadır.

Tablo 1'de verilen bütünsel dereceli puanlama anahtarı kullanılarak elde edilen verilerde b x g x p desenine ait G çalışması sonucunda kestirilen varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde, en çok birey (b) ana etkisine ait varyans bileşeninin toplam varyansın % 76.0'sını açıkladığı, en az ise birey x görev ortak etkileşiminin % 0.0 değeri ile toplam varyansa katkı sağlamadığı görülmektedir. Her iki görevin benzer güçlükte olduğu, $\boldsymbol{\sigma_g^2}$(0.243) varyans bileşen değeri ve % 1.2 toplam varyansı açıklama yüzdesiyle görülmektedir. Puanlayıcıların bireylerin performansını değerlendirirken birbirlerine göre daha cömert (yumuşak) olmadığını, $\boldsymbol{\sigma_p^2}$(0.310) varyans bileşen değeri ve % 1.6 toplam varyansı açıklama yüzdesiyle görülmektedir. Farklı puanlayıcıların benzer puanlamalar yaptığı anlaşılmaktadır. Bireylerin durumlarının bir puanlayıcıdan diğerine kısmen farklılık gösterdiği, $\boldsymbol{\sigma_{bp}^2}$(2.191) varyans bileşen değeri ve % 11.2 toplam varyansı açıklama oranıyla görülmektedir. Ölçmenin nesnesi birey ve b x p etkileşiminden sonra en büyük değişkenlik kaynağının tesadüfi hata olarak adlandırılan değişkenlik kaynağı olduğu, $\boldsymbol{\sigma_{bgp,e}^2}$(1.901) varyans bileşen değeri ve % 9.8 toplam varyansı açıklama yüzdesiyle anlaşılmaktadır.

Tablo 1'de verilen b x g x p x a desenine ait G çalışması sonucunda kestirilen varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde, en çok birey (b) ana etkisine ait varyans bileşeninin toplam varyansın % 72.5'ini açıkladığı, en az ise birey x görev, görev x puanlayıcı, puanlayıcı x anahtar ortak etkileşiminin % 0.0 değeri ile toplam varyansa katkı sağlamadığı görülmektedir. G kuramında kestirilen varyans bileşeninin negatif değerli olması durumunda, Cronbach ve diğerleri varyans bileşenlerini hesaplamak için negatif olan varyans bileşeninin sıfır alınmasını önermiştir (Brennan, 2001a). Yurt içinde yapılan araştırmalarda da, negatif çıkan varyans bileşeni yerine sıfır değerinin yazıldığı görülmektedir (Atılgan, 2004; Taşdelen ve diğ., 2010). Bu yüzden birey x görev ortak etkisinin varyans bileşeni σ_{bg}^2(-0.136) iken Tablo 1'de σ_{bg}^2(0.00) kabul edilmiştir. Yine görev x puanlayıcı ortak etkisinin varyans bileşeni σ_{gp}^2(-0.001) ve puanlayıcı x anahtar ortak etkisinin varyans bileşeni σ_{pa}^2(-0.004) olup negatifken Tablo 1'de σ_{gp}^2(0.00) ve σ_{pa}^2(0.00) kabul edilmiştir.

K Çalışması Bulguları

Analitik ve bütünsel dereceli puanlama anahtarından elde edilen puanlarda tüm değişkenlerin çaprazlandığı b x g x p ve anahtar türü değişkenlik kaynağı olarak alındığı b x g x p x a desenlerinde birey ölçmenin nesnesi (object of measurement) olarak belirlenip görev ve puanlayıcı sayılarının artırılıp azaltılmasıyla yapılan senaryolar için kestirilen genellenebilirlik katsayısı (G), Phi katsayısı, bağıl hata varyansı ve mutlak hata varyanslarına ait değerler Tablo 2'de verilmiştir.

Tablo 2. b x g x p ve b x g x p x a Desenlerinde Puanlayıcı ve Görev Sayılarının Artırılıp Azaltılmasıyla Yapılan Karar Çalışması (K) Sonuçları

		Analitik dereceli puanlama anahtarı **b x g x p**		*Bütünsel dereceli puanlama anahtarı* **b x g x p**		**b x g x p x a**	
n_g	**n_p**	**G**	**Phi**	**G**	**Phi**	**G**	**Phi**
1	2	0.897	0.889	0.878	0.857	0.871	0.857
1	4	0.946	0.938	0.935	0.916	0.882	0.872
1	6	0.963	0.955	0.955	0.937	0.886	0.877
2	2	0.913	0.908	0.904	0.888	0.884	0.873
2	3	0.940	0.935	0.934	0.920	0.889	0.880
2	**4**	**0.954**	**0.950**	**0.949**	**0.937**	**0.892**	**0.884**
2	5	0.965	0.962	0.959	0.947	0.893	0.886
2	6	0.969	0.965	0.965	0.954	0.894	0.888
3	2	0.918	0.914	0.913	0.899	0.888	0.878
3	4	0.957	0.954	0.954	0.944	0.895	0.888
3	6	0.971	0.968	0.969	0.960	0.897	0.892

ng: görev sayısı, np: puanlayıcı sayısı

Araştırmada kullanılan b x g x p deseninde 132 bireyden her birinin, dört puanlayıcı tarafından, iki görev doğrultusunda analitik dereceli puanlama anahtarıyla puanlanmasıyla elde edilen puanların G katsayısı 0.954, phi katsayısı ise 0.950 olarak kestirilmiştir. Tablo 2 incelendiğinde; bu dereceli puanlama anahtarı için b x g x p deseninde puanlayıcı sayısını iki arttırıp azaltmanın G ve phi katsayısı üzerindeki etkisinin, görev sayısını bir arttırıp azaltmadan daha fazla olduğu görülmektedir. Ayrıca görev ve puanlayıcı sayısını arttırmanın G ve phi katsayısını arttırdığı, azaltmanın ise G ve Phi katsayısını azalttığı görülmektedir. Tablo 2'de de görüleceği gibi puanlayıcı ve görev sayısının en az olduğu durumda ($n_p = 2$, $n_g = 1$) G katsayısı 0.897 ve phi katsayısı 0.889 olup, yapılan karar çalışmasında en düşük güvenirlik değerlerini almaktadır. Puanlayıcı ve görev sayısının en az olduğu durumda ($n_p = 2$, $n_g = 1$) G katsayısı 0.897 ve phi katsayısı 0.889 olup, yapılan karar çalışmasında en düşük güvenirlik değerlerini almaktadır. Puanlayıcı ve görev sayısının en fazla olduğu durumda ($n_p = 6$, $n_g = 3$) ise, G katsayısı 0.971 ve phi katsayısı 0.968 olup, yapılan karar çalışmasında en yüksek güvenirlik değerlerini almaktadır.

Bütünsel dereceli puanlama anahtarından elde edilen puanlarda b x g x p deseninde ise; G katsayısı 0.949, Phi katsayısı ise 0.937 olarak kestirilmiştir. Tablo 2 incelendiğinde; puanlayıcı sayısını iki arttırıp azaltmanın G ve phi katsayısı üzerindeki etkisinin, görev sayısını bir arttırıp azaltmanın etkisinden daha fazla olduğu görülmektedir. Bunun dışında Tablo 2 incelendiğinde görev ve puanlayıcı sayısını arttırmanın G ve phi katsayısını arttırdığı, azaltmanın ise G ve phi katsayısını azalttığı görülmektedir. Puanlayıcı ve görev sayısının en az olduğu durumda ($n_p = 2$, $n_g = 1$) G katsayısı 0.878 ve phi katsayısı 0.857 olup, yapılan karar çalışmasında en düşük güvenirlik değerlerini almaktadır. Puanlayıcı ve görev sayısının en fazla olduğu durumda ($n_p = 6$, $n_g = 3$) ise, G katsayısı 0.969 ve phi katsayısı 0.960 olup, yapılan karar çalışmasında en yüksek güvenirlik değerlerini almaktadır.

Araştırmada 132 bireyden her birinin dört puanlayıcı tarafından iki görev doğrultusunda analitik ve bütünsel dereceli puanlama anahtarıyla puanlanmasıyla elde edilen puanların b x g x p x a deseninde G katsayısı 0.892, phi katsayısı ise 0.884 olarak kestirilmiştir. Tablo 2 incelendiğinde yine görev ve puanlayıcı sayısını arttırmanın G ve phi katsayısını arttırdığı, azaltmanın ise G ve phi katsayısını azalttığı görülmektedir. Puanlayıcı ve görev sayısının en az olduğu durumda ($n_p = 2$, $n_g = 1$) G katsayısı 0.871 ve phi katsayısı 0.857 olup, yapılan karar çalışmasında en düşük güvenirlik değerlerini almaktadır. Puanlayıcı ve görev sayısının en fazla olduğu durumda ($n_p = 6$, $n_g = 3$) ise, G katsayısı 0.897 ve phi katsayısı 0.892 olup, yapılan karar çalışmasında en yüksek güvenirlik değerlerini almaktadır.

Her İki Desenden Elde Edilen Bulguların Karşılaştırılması

Tablo 1'deki b x g x p ve b x g x p x a desenlerine ait G çalışması sonucunda toplam varyansı açıklama yüzdeleri incelendiğinde, en fazla birey (b) ana etkisinin varyans açıklama yüzdesine sahip olduğu, analitik dereceli puanlama anahtarından elde edilen veriler için % 80.6 toplam varyansı açıklama yüzdesi, bütünsel dereceli puanlama anahtarından elde edilen veriler için %76 toplam varyansı açıklama yüzdesi, anahtar değişkenlik kaynağı olduğu durumda %72,5 toplam varyansı açıklama yüzdesiyle görülmektedir. Böylece değişkenlik kaynağı sayısı arttıkça ölçmenin nesnesi (objesi) bireyin toplam varyansı açıklama yüzdesinin azaldığı görülmektedir. Ayrıca bu çalışmada analitik dereceli puanlama anahtarı kullanmanın bütünsel dereceli puanlama anahtarı kullanmaya oranla ölçmenin nesnesinin (objesi) varyans açıklama yüzdesini arttırdığı görülmektedir.

Kullanılan her üç desende birey x görev x puanlayıcı ortak etkisi incelendiğinde analitik dereceli puanlama anahtarıyla elde edilen verilerde ölçme hatasından kaynaklı artık değişkenlik kaynağı; % 6.1 toplam varyansı açıklama yüzdesine, bütünsel dereceli puanlama anahtarıyla elde edilen verilerde % 9.8 toplam varyansı açıklama yüzdesine, anahtar değişkenlik kaynağı olduğu durumda b x g x p x a deseninde ise % 0.1 toplam varyansı açıklama yüzdesine sahip olduğu görülmektedir. Bu bulguların elde edilmesinde b x g x p x a deseninde anahtar türünün değişkenlik kaynağı olmasıyla diğer değişkenlik kaynaklarının toplam varyansa etkisini paylaşmasının etkili olabileceği düşünülmektedir. Bireylerin durumlarının bir anahtardan diğerine kısmen farklılık gösterdiği b x g x p x a deseninde % 15.1 toplam varyansı açıklama yüzdesiyle görülmektedir.

Karar çalışması bulguları incelendiğinde ise analitik dereceli puanlama anahtarlarından elde edilen verilerin kullanıldığı G ve phi katsayıları sırasıyla katsayısı 0.954, Phi katsayısı ise 0.950 G katsayısı 0.949, phi katsayısı ise 0.937 olarak kestirilmiştir. b x g x p x a deseninde G katsayısı 0.892, phi katsayısı ise 0.884 olarak kestirilerek b x g x p desenine göre daha düşük değerler elde edilmişir.

Tartışma, Sonuç ve Öneriler

Analitik dereceli puanlama anahtarı ile elde edilen puanlar üzerine b x g x p tümüyle çaprazlanmış desenle G çalışması yapıldığında, bütünsel puanlama anahtarıyla elde edilen puanlara oranla daha güvenilir sonuçlara ulaşılmıştır. Tesadüfi hata olarak da adlandırılan artık etki analitik dereceli puanlama anahtarından elde edilen puanlarda daha düşük değerde çıkarken, ölçmenin amacı olan bireylerin problem çözme yeteneklerindeki farklılıktan kaynaklanan etki ise daha yüksek çıkmıştır. Bu sonuçlar alanyazında analitik dereceli puanlama anahtarından elde edilen puanların, bütünsel dereceli puanlama anahtarından elde edilen puanlara göre klasik test kuramında göreceli olarak daha yüksek güvenirlik gösterdiği birçok çalışmayla örtüşmekte (Bauer, 1981; Follman ve Anderson, 1967; Jonsson ve Svingby, 2007) ve analitik dereceli puanlama anahtarından elde edilen puanların, bütünsel dereceli puanlama anahtarından elde edilen puanlara göre önemli derecede güvenilir olduğu sonucuna ulaşan çalışmaların bulgularıyla (Boring, 2002; Klein et. al, 1998) kısmen örtüşmektedir.

Kullanılan desende (b x g x p) K çalışması yapıldığında ise yine analitik dereceli puanlama anahtarından elde edilen puanların, bütünsel dereceli puanlama anahtarından elde edilen puanlara oranla daha yüksek G ve phi katsayılarına sahip olduğu görülmüştür.

Her iki puanlama anahtarında da görev ve puanlayıcı sayılarının arttırılması durumlarında elde edilen G ve phi katsayılarının da kısmen arttığı, ancak puanlayıcı sayısını arttırmanın, görev sayısını arttırmaya göre; katsayıları arttırmada az miktarda da olsa, daha fazla etkisi olduğu görülmüştür.

Diğer kullanılan desende (b x g x p) olduğu gibi b x g x p x a deseninde de görev ve puanlayıcı sayılarının arttırılması durumlarında elde edilen G ve phi katsayılarının da kısmen arttığı, ancak b x g x p desenindeki karar çalışmalarından elde edilen bulguların aksine görev sayısını arttırmanın, puanlayıcı sayısını arttırmaya göre; katsayıları arttırmada az miktarda da olsa, daha fazla etkisi olduğu görülmüştür. Ayrıca kullanılan desende değişkenlik kaynağı sayısını arttırmanın göreceli olarak G ve phi katsayısını düşürdüğü b x g x p ve b x g x p x a deseninlerinde yapılan karar çalışmalarından anlaşılmaktadır. Bu durumda toplam varyansın daha fazla değişkenlik kaynakları arasında paylaşılmasından kaynaklı olabileceği düşünülmektedir.

Sonuç olarak uygulayıcılar performansa dayalı durum belirlemede puanlayıcı güvenirliğini arttırmak için ölçmenin amacını ve öğrencilerin düşünme stilleri göz önüne alarak davranış tanıma kodlarıyla göreve özgü analitik dereceli puanlama anahtarı kullanabilir. Araştırmacılar daha büyük araştırma grubu, puanlayıcı ve görevle çalışarak genellenebilirlik kuramında farklı değişkenlik kaynaklarının ele alındığı desenlerle güvenirlik ve genellenebilirlik çalışması yürütebilir. Performansa dayalı durum belirlemede puanlayıcılar arası tutarlılık düşük çıktığı durumlarda bu durumun nedenini sorgulamak için nitel çalışmalarla güvenirlik çalışmaları beraber yürütülebilir.

Kaynakça

Al-Mahroos, F. (2009). Construct validity and generalizability of pediatrics clerkship evaluation at a problem-based medical school, Bahrain. *Evaluation & the Health Professions, 32*(2), 165-183.

American Educational Research Association, American Psychological Association ve National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Arce-Ferrer, A. J. ve Castillo, I. B. (2007). Investigating postgraduate college admission interviews: generalizability theory reliability and incremental predictive validity. *Journal of Hispanic Higher Education, 6(2)*, 118-134.

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Ankara.

Bauer, B. A. (1981). *A study of the reliabilities and cost-efficiencies of three methods of assessment for writing ability. (*ERIC Document Reproduction Service No. ED 216357).

Boring, R. L. (2002). *Human and computerized essay assessment: a comparative analysis of holistic, analytic and latent semantic methods*. Unpublished thesis, Department of Psychology, New Mexico State University, Las Cruses, New Mexico.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2001b). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38(*4), 295-317.

Crocker, L. M. ve Algina, L. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Christ, T. J., Tillman C., Chafouleas, S. M. ve Boice C. H. (2010). Direct behavior rating (DBR): generalizability and dependability across raters and observations. *Educational and Psychological Measurement, 70(5),* 825-843.

Çakıcı, D. (2011). *Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlığın karşılaştırılması.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması.* Yayımlanmamış doktora tezi, Ankara Üniversitesi, Ankara.

Follman, J. C. ve Anderson, J. A. (1967). An investigation of reliability of five procedures for grading english themes. *Research in the Teaching of English, 1(*2), 190-200.

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve rasch modeli üzerine bir araştırma.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Ankara.

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Eğitim ve Bilim, 34*(154), 93-103.

Güler, N. (2011). Rastgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramı'na göre güvenirliğin incelenmesi. *Eğitim ve Bilim, 36*(162), 225-234.

Hoyt, W. T. ve Melby, J. N. (1999). Dependability of measurement in counseling psychology: an introduction to generalizability theory. *The Counseling Psychologist,* 27(3), 325-352.

Jarjoura, D., Early, L. ve Androulakakis, V. (2004). A multivariate generalizability model for clinical skills assessments. *Educational and Psychological Measurement, 64(1),* 22-39.

Jonsson, A. ve Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review, 2(2),* 130-144.

Kaya, G. (2011). *Genellenebilirlik kuramının doldurma kavram haritası değerlendirme çalışmasına uygulanması.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Kozaki, Y. (2004). Using GENOVA and SPSS to set multiple standards on from japanese into english performance assessment for certification in medical translation. *Language Testing, 21*(1), 1-27.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K. ve Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11(2)*, 121-137.

Lane, S. ve Stone, C. A. (2006). *Performance assessment.* Brennan, R.L. (Ed.) Educational Measurement (4th ed.). 387-431. Westport, CT: Praeger Puublishers.

Linn, R. L. ve Miller D. M. (2004). *Measurement and assesment in teaching.* (9th edition). Upper Saddle River: Printice-Hall Inc.

Martin J. Bergee. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55(4)*, 344-358.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7(*25). 1 Kasım 2011 tarihinde http://PAREonline.net/getvn.asp?v=7&n=25 adresinden erişildi.

Moskal, B. M. (2000). Scoring rubrics: what, when and how? *Practical Assessment, Research & Evaluation, 7(*3). 30 Ekim 2011 tarihinde http://PAREonline.net/getvn.asp?v=7&n=3 adresinden erişildi.

Mushquash, C. ve O'Connor, B. P. (2006). SPSS and SAS programs for Generalizability Theory analysis. *Behavior Research Methods, 38*(3), 542-547.

Nalbantoğlu, F. (2009). *Performans ölçümlerinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Nalbantoğlu, F. (2012*). Genellenebilirlik kuramında dengelenmiş ve dengelenmemiş desenlerin karşılaştırılması -ıntramuskuler enjeksiyon yapma istasyon verileri üzerine bir uygulama-.* Yayımlanmamış doktora tezi, Ankara Üniversitesi, Ankara.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramı'na göre karşılaştırılması.* Yayınlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Polya, G. (1973). *How to sove it? A new aspect of mathematical method.* Second edition. Princeton University press. New Jersey.

Popham, J. W. (1997). What's wrong and what's right with rubric. *Educational Leadership, 55*(2), 72-75.

Priestley, M. (1982). *Performance assessment in education and training: alternative techniques.* First edition. Educational Technology Publications. New Jersey.

Shavelson, R. J. ve Webb, N. M. (1991). *Generalizability theory: a primer.* Sage Publications, USA.

Taşdelen, G., Kelecioğlu, H. ve Güler, N. (2010). Nedelsky ve angoff standart belirleme yöntemleri ile elde edilen kesme puanlarının genellenebilirlik kuramı ile karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(1), 22-28.

Tindal, G., Yovanoff, P. ve Geller, J. P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *The Journal of Special Education, 44*(1), 3-17.

Van Hooft, E. J. A., Born, M., Taris, T. W. ve Van Der Flier, H. (2006). The cross-cultural generalizability of the theory of planned behavior: a study on job seeking in the Netherlands. *Journal of Cross-Cultural Psychology, 37*(2), 127-135.

Yelboğa, A. (2007). *Klasik test kuramı ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi*. Yayımlanmamış doktora tezi, Ankara Üniversitesi, Ankara.

Yelboğa, A. (2012). Genellenebilirlik kuramına göre iş performans ölçeklerinde güvenirlik. *Eğitim ve Bilim, 37*(163), 157-164.